

- 1 -

cDNA MICROARRAY DATA CORRECTION SYSTEM,  
METHOD, PROGRAM, AND MEMORY MEDIUM

5

BACKGROUND OF THE INVENTION

The present invention relates to a data correction system, method, program, and memory medium for cDNA microarray data based on a mathematical model, and more particularly to a cDNA microarray data correction system, method, program, and memory medium enabling global normalization and local normalization and further enabling a correction of a distortion of measurements caused by a difference in sensitivity between fluorescent dyes.

15

Related Background Art

The genome research is developing from an individual gene structural analysis into a systematic gene functional analysis at present. Experiments using complementary DNA (cDNA) microarray is greatly expected to display effectiveness due to its ability to quantify an expression intensity of a lot of genes simultaneously for a functional analysis of genes whose functions are unknown or genes in the mass.

20

The purpose of an experiment using the cDNA microarray in a two-color fluorescence technique is to detect a difference in gene expression between two types of cells. An outline of the cDNA microarray in the two-color fluorescence technique is then described here. First, a

25

- 2 -

lot of gene sets of cDNA are densely fixed as a reference probe in an array on a slide glass (microarray).

Subsequently, mRNAs sampled from two types of samples under different conditions, cell 1 and cell 2 (for example, a normal cell and a cancer cell) are labeled with fluorescent dyes having different wavelengths for a synthesis of a target cDNA. Then, those mixed in equal proportions are used for competitive hybridization with the reference probe cDNA fixed to the microarray. After the hybridization, the intensities of the fluorescent dyes are measured using a scanner. The fluorescent dye on the cell 1 and the fluorescent dye on the cell 2 are read in channel 1 and channel 2, respectively, and they are considered to be gene expression intensity data of the respective cells (microarray data).

Thus, the process of achieving the microarray data is complicated and requires advanced experimental techniques. Thereby various experimental errors are anticipated in various stages of the experiment. Therefore, an analysis of a distribution of gene expression intensities and experimental errors are important problems in order to extract truly biologically significant data from the microarray data.

Regarding the distribution of gene expression intensities, for example, with reference to a document 1 (Journal of Computational Biology Vol. 8, pp. 37-52), Newton et al. considered the statistical property of a gene expression intensity ratio (ratio of gene expression intensity data between channel 1 and channel 2), assuming

- 3 -

gamma distribution functions for gene expression intensities.

In addition, for observed gene expression intensity data, for example, with reference to a document 2 (Proceeding of the National Academy of Sciences Vol. 97, No 5 18, pp. 9834-9839), Lee et al. applied a mixture normal distribution as represented by the following EQ1 to the statistical consideration of the gene expression intensity data, on the assumption that true gene expression 10 intensities can be separated into two levels and that random errors exist.

$$f(x) = p\phi(x - \mu_1 | \sigma_1^2) + (1 - p)\phi(x - \mu_2 | \sigma_2^2) \quad (1)$$

In the above,  $x$  indicates gene expression intensity data such as a fluorescence intensity obtained 15 using a scanner,  $\phi(x - \mu_1 | \sigma_1^2)$  in the first term of the right-hand side indicates a density function of a normal distribution of average  $\mu_1$  and variance  $\sigma_1^2$  in a gene expression state,  $\phi(x - \mu_2 | \sigma_2^2)$  in the second term of the right-hand side indicates a density function of a normal 20 distribution of average  $\mu_2$  and variance  $\sigma_2^2$  in a gene non-expression state, and  $p$  is a population parameter indicating their mixing rate.

Regarding the analysis of experimental errors, some methods of removing systematic errors, namely, 25 normalization methods have been suggested. There are main two types of normalization; global normalization intended

- 4 -

for all spots on an array and local standardization intended for spots in units of a subset (for example, in units of a grid). Regarding the global normalization, for example, with reference to a document 3 (Journal of Biomedical Optics Vol. 2, pp. 364-374), Chen et al. corrected measurements obtained in channel 1 and channel 2 assuming that medians of gene expression intensities of two cells are equal. Regarding the local normalization, for example, with reference to a documents 4 (Dudoit et. al, 2000. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical-Report #578 2.), 5(Nucleic Acids Research, 2000, Vol.28, No. 10), and 6(Nucleic Acids Research, 2002, Vol.30, No. 4), Dudoit, Shuchhardt, and Yang thought that systematic errors are caused by differences in a spot position on a slide glass or in sensitivity between two types of fluorescent dyes and suggested a method for removing them.

#### SUMMARY OF THE INVENTION

A problem in the conventional technologies in the above is that the microarray data lacks in reproducibility with a tendency to be unstable and thereby its precision or efficiency is considered to be low. It is because true signals on the gene expression are not fully separated from experimental errors. As a factor behind it, it is attributed to the fact that the gene expression intensities could have various levels depending on the gene. If so, the above model represented by EQ1 is apparently too

- 5 -

simplified.

It is an object of the present invention to provide a comprehensive normalization method and system for correcting global and local distortions with high precision and further correcting measurement errors caused by a difference in sensitivity between fluorescent dyes by assuming more accountable mathematical model of gene expression intensity data on a microarray.

In accordance with an aspect of the present invention, there is provided a cDNA microarray data correction system, comprising an input device for inputting gene expression intensity data such as a fluorescence intensity, a data analyzer operating with program controls, and an output device. It is assumed that the gene expression intensity data to be input is previously adjusted, considering flag information indicating a removal of background noise of each spot and reliability of the spot.

The data analyzer has the following three continuous processes. A data standardization unit for a first process inputs gene expression intensity data from the input device, standardizes the gene expression intensity data by using grid-by-grid order statistics on the assumption that most genes are in a non-expression state, and outputs the standardized gene expression intensity data.

A spot-position-based correction unit for a second process inputs the standardized gene expression intensity data, estimates a distortion depending on the spot position

- 6 -

on a grid by grid basis by a nonparametric smoothing method, and outputs gene expression intensity data whose distortion depending on the spot position has been corrected.

5 An S-D-plot-based correction unit for a third process performs an S-D transformation, which is a variation of an MA transformation (for information about the MA transformation and an MA plot, refer to the above nonpatent document 6), for the gene expression intensity data corrected up to the second process, estimates a  
10 potential distortion caused by a difference in sensitivity between the fluorescent dyes in the gene expression intensity data by the nonparametric smoothing method, and outputs the gene expression intensity data whose distortion caused by the difference in sensitivity between the  
15 fluorescent dyes has been corrected to the output device.

This system further comprises an S-D transformation unit for quantifying the distortion of the gene expression intensity data in an arbitrary stage and for visualizing it on the S-D plot.

20 By using the constitution to correct the gene expression intensity data, the object of the present invention can be achieved.

#### BRIEF DESCRIPTION OF THE DRAWINGS

25 Fig. 1 is a diagram showing a microarray structure according to the present invention;

Fig. 2 is a block diagram showing a configuration of a first embodiment of the present invention;

Fig. 3 is a flowchart showing an operation of the

- 7 -

first embodiment of the present invention;

Fig. 4 is a block diagram showing a configuration of a second embodiment of the present invention;

5 Fig. 5 is a diagram showing gene expression intensities of original data obtained in channel 1;

Fig. 6 is a diagram showing gene expression intensities of original data obtained in channel 2;

10 Fig. 7 is a diagram showing gene expression intensities of original data (a first grid to a fourth grid) obtained in the channel 1;

Fig. 8 is a diagram showing gene expression intensities of original data (a first grid to a fourth grid) obtained in the channel 2;

15 Fig. 9 is an S-D plot to the original data;

Fig. 10 is a diagram showing gene expression intensities of the original data in the channel 1;

Fig. 11 is a diagram showing gene expression intensities after a first process in the channel 1;

20 Fig. 12 is a diagram showing gene expression intensities after a second process in the channel 1;

Fig. 13 is a diagram showing gene expression intensities after a third process in the channel 1;

Fig. 14 is a diagram showing gene expression intensities of the original data in the channel 2;

25 Fig. 15 is a diagram showing gene expression intensities after the first process in the channel 2;

Fig. 16 is a diagram showing gene expression intensities after the second process in the channel 2;

Fig. 17 is a diagram showing gene expression

- 8 -

intensities after the third process in the channel 2;

Fig. 18 is an S-D plot to the original data;

Fig. 19 is an S-D plot after the first process;

Fig. 20 is an S-D plot after the second process;

5. and

Fig. 21 is an S-D plot after the third process.

#### DESCRIPTION OF THE PREFERRED EMBODIMENTS

First, a microarray structure in the present  
10 invention will be described hereinafter. Referring to Fig.  
1, there are cDNA spots on a slide glass having K grids by  
I×J spots per grid, that is, K×I×J spots in total. In this  
condition, it is assumed that  $y_{ij}^k(c)$ ,  $c=1, 2$  is a  
fluorescence intensity obtained in channel c ( $c = 1, 2$ ) for  
15 the cDNA spots on the coordinates (i, j) in grid k.

Subsequently, the following two assumptions are  
provided.

Supposing that a probability of gene expression is  
lower than 0.5, it is assumed that the fluorescence  
20 intensity  $y_{ij}^k(c)$  detected at more than half of the spots  
within each grid indicates a background noise or a  
systematic error (Assumption 1).

Supposing that  $L_k(c)$  and  $M_k(c)$  indicate 25 and 50  
percent points of the fluorescence intensity  $y_{ij}^k(c)$   
25 obtained in the channel c in the grid k, then it is assumed  
that  $L_k(c)$  and  $M_k(c) - L_k(c)$  are common to all grids and  
channels on condition that most genes are in a non-  
expression state and that the distribution under 50 percent  
point of the fluorescence intensity is common to all grids

- 9 -

and channels (Assumption 2).

Subsequently, based on the above assumptions, the first embodiment of the present invention will be described in detail by referring to appended drawings. Referring to  
5 Fig. 2, there is shown the first embodiment of the present invention comprising an input device 1 for inputting gene expression intensity data such as a fluorescence intensity, a data analyzer 2 operating with program controls, and an output device 3 such as a display unit or a printer. The  
10 data analyzer includes a data standardization unit 21, a spot-position-based correction unit 22, and an S-D-plot-based correction unit 23.

The data standardization unit 21 standardizes gene expression intensity data by using grid-by-grid order  
15 statistics for given gene expression intensity data and then transmits it to the spot-position-based correction unit 22 and an S-D transformation unit 24.

The spot-position-based correction unit 22 estimates a distortion depending on the spot position on a  
20 grid by grid basis by a nonparametric smoothing method for the standardized gene expression intensity data transmitted from the data standardization unit 21 and then transmits corrected gene expression intensity data to the S-D-plot-based correction unit 23 and the S-D transformation unit 24.

25 The S-D-plot-based correction unit 23 performs an S-D transformation for the corrected gene expression intensity data transmitted from the spot-position-based correction unit 22, corrects a distortion caused by a difference in sensitivity between fluorescent dyes by the

- 10 -

nonparametric smoothing method, and outputs the gene expression intensity data to the output device 3.

The S-D transformation unit 24 performs an S-D transformation for the gene expression intensity data transmitted from the spot-position-based correction unit 22 and then transmits it to the output device 3.

Subsequently, the embodiment will be described in detail by referring to Fig. 2 and Fig. 3. The gene expression intensity data such as fluorescence intensity input from the input device 1 is transmitted to the data standardization unit 21. The data standardization unit 21 standardizes the expression intensity data it has received by using grid-by-grid order statistics as represented by the following EQ2 (step A1 in Fig. 3).

$$w_{ij}^k(c) = \frac{y_{ij}^k(c) - L_k(c)}{M_k(c) - L_k(c)},$$

$$c = 1, 2, i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K. \quad (2)$$

It is determined whether the gene expression intensity data  $y_{ij}^k(c)$  of all spots obtained in the two channels have been standardized. This operation is continued until the standardization of the gene expression intensity data ( $2 \times I \times J \times K$  pieces) of all spots is completed (step A2).

For the gene expression intensity data  $w_{ij}^k(c)$  standardized by the data standardization unit 21, it is assumed that  $z_{ij}^k(c)$  is a fluorescence intensity reflecting a true expression intensity (hereinafter, referred to as a true expression fluorescence intensity) and that  $\xi_{ij}^k(c)$  is a spot-position-dependent distortion on coordinates  $(i, j)$  of

- 11 -

the grid  $k$ . In this condition, it is assumed that gene expression intensity data  $w_{ij}^k(c)$  is represented by a sum of the true expression intensity  $z_{ij}^k(c)$  and the spot-position-dependent distortion  $\xi_{ij}^k(c)$ , as represented by the following EQ3.

$$w_{ij}^k(c) = z_{ij}^k(c) + \xi_{ij}^k(c) + \varepsilon_{ij}^k(c), \varepsilon_{ij}^k(c) \sim N(0, \sigma_k(c)^2), c = 1, 2. \quad (3)$$

In the above,  $\varepsilon_{ij}^k(c)$  is assumed a random noise.

The spot-position-based correction unit 22 describes the spot-position-dependent distortion  $\xi_{ij}^k(c)$  by means of a nonparametric regression model represented by a regression relation of distortions with an x-axis, a y-axis, and an interaction of the two axes as represented by the following EQ4 and estimates the spot-position-dependent distortion  $\xi_{ij}^k(c)$  by using the nonparametric smoothing method as represented by the following EQ5.

$$\xi_{ij}^k(c) = \alpha_k^{(c)}(i) + \beta_k^{(c)}(j) + \gamma_k^{(c)}((i - m_i)(j - m_j)), c = 1, 2, i = 1, \dots, I, j = 1, \dots, J,$$

$$\sum_i \alpha_k^{(c)}(i) = 0, \sum_j \beta_k^{(c)}(j) = 0, \sum_i \sum_j \gamma_k^{(c)}((i - m_i)(j - m_j)) = 0. \quad (4)$$

$$\hat{\xi}_{ij}^k(c) = \hat{\alpha}_k^{(c)}(i) + \hat{\beta}_k^{(c)}(j) + \hat{\gamma}_k^{(c)}((i - m_i)(j - m_j)), c = 1, 2, i = 1, \dots, I, j = 1, \dots, J. \quad (5)$$

In the above,  $m_i = \lfloor I/2 \rfloor$ ,  $m_j = \lfloor J/2 \rfloor$  is assumed.  $\lfloor \alpha \rfloor$  is assumed to be the minimum integer equal to or more than  $\alpha$ .

The spot-position-based correction unit 22 corrects the estimated spot-position-dependent distortion  $\hat{\xi}_{ij}^k(c)$  for the gene expression intensity data  $w_{ij}^k(c)$  standardized by the data standardization unit 21 (step A3) as represented by the following EQ6.

- 12 -

$$\hat{z}_{ij}^k(c) = w_{ij}^k(c) - \hat{\xi}_{ij}^k(c) \quad (6)$$

It is determined whether the spot-position-dependent distortion  $\hat{\xi}_{ij}^k(c)$  has been corrected for the gene expression intensity data  $w_{ij}^k(c)$  of all spots standardized by the data standardization unit 21. This operation is continued until the correction is completed for the gene expression intensity data ( $2 \times I \times J \times K$  pieces) of all spots (step A4).

The S-D-plot-based correction unit 23 performs the S-D transformation for the true gene expression intensity data  $\hat{z}_{ij}^k(c)$  corrected by the spot-position-based correction unit 22 as represented by the following EQ7.

$$\begin{aligned} u_{ij}^k &= \hat{z}_{ij}^k(1) + \hat{z}_{ij}^k(2) \\ v_{ij}^k &= \hat{z}_{ij}^k(1) - \hat{z}_{ij}^k(2) \end{aligned} \quad (7)$$

Furthermore, with a description of a nonparametric regression model as represented by the following EQ8, a measurement error caused by a difference in sensitivity between the fluorescent dyes is corrected after estimating it by the nonparametric smoothing method as represented by the following EQ9 and EQ10 (step A5).

$$v_{ij}^k = \phi(u_{ij}^k) + \varepsilon_{ij}^k, \quad \varepsilon_{ij}^k \sim N(0, \sigma^2) \quad (8)$$

$$\eta_{ij}^k = v_{ij}^k - \hat{\phi}(u_{ij}^k) \quad (9)$$

- 13 -

$$\begin{aligned}\hat{y}_{ij}^k(1) &= \frac{1}{2} (u_{ij}^k + \eta_{ij}^k) \\ \hat{y}_{ij}^k(2) &= \frac{1}{2} (u_{ij}^k - \eta_{ij}^k)\end{aligned}\quad (10)$$

It is determined whether the correction with the S-D plot has been performed for the true gene expression intensity data  $\hat{z}_{ij}^k(c)$  corrected by the spot-position-based correction unit 22. This operation is continued until the  
5 correction is completed for the true gene expression intensity data ( $2 \times I \times J \times K$  pieces) of all spots (step A6).

After a completion of the steps A2 and A4 in Fig. 3, the gene expression intensity data is transmitted to the  
10 output device 3 via the S-D transformation unit 24, by which the distortion of the gene expression intensity data can be visualized by the S-D plot.

Subsequently, effects of the embodiment will be described below. In the embodiment, the standardization  
15 has been made by combining standardization using order statistics over the grids (global standardization) and the correction of a distortion depending on the spot position within a grid (local standardization). Thereby, it becomes possible to correct a systematic error caused by deviation  
20 of the gene expression intensities among the grids and a distortion depending on the spot position within an individual grid at a time. Furthermore, in the correction with the S-D plot, the measurement error caused by a difference in sensitivity between fluorescent dyes can be  
25 corrected by using a sum and a difference of the expression intensity data.

- 14 -

A second embodiment of the present invention will now be described in detail hereinafter by referring to appended drawings. Referring to Fig. 4, there is shown the second embodiment of the present invention comprising an input device, a data analyzer, and an output device similarly to the first embodiment, and further comprising a memory medium 4 where a data analysis program is recorded. The memory medium 4 can be either transportable or of stationary type, and can be a magnetic disk, a semiconductor memory, a CD-ROM, or any other memory medium.

In addition, it is also possible to previously store a computer program capable of executing the method of the present invention in a recording device of a computer connected to a network and to transfer the program to another computer via the network. A medium for providing the computer program for executing this algorithm can be distributed as a medium whose data can be read out to computers in various formats, and it is not limited to a specific type of mediums. The data analysis program is read from the memory medium 4 to a data analyzer 5 to control operations of the data analyzer 5, thereby executing the same processing as in the data analyzer 2 of the first embodiment for the data file input from the input device 1.

The embodiment of the present invention will be described hereinafter. Data for the example is obtained from an experiment for a comparison between two different types of cancer cells (cell A and cell B) in the gene expression condition.

- 15 -

The following is a result of checking gene expression patterns in 48 grids on a single chip with 441 (21×21) spots per grid, namely, 21,168 spots in total.

Referring to Fig. 5 and Fig. 7, there are shown gene expression intensities of the cell A of original data obtained in channel 1. Referring to Fig. 6 and Fig. 8, there are shown gene expression intensities of the cell B of original data obtained in channel 2. These graphs show plotting of logarithmic values of the gene expression intensities of the spot positions on the microarray. Fig. 7 and Fig. 8 are enlarged views of the first to fourth grids. As is shown by Fig. 5 to Fig. 8, we can observe a systematic distortion periodically repeated in gene expression intensities on a grid by grid basis. Since the gene spots are arranged at random on the microarray, the distortion is thought to be an experimental error.

Referring to Fig. 9, there is shown an S-D plot of the distortion. The abscissa indicates a sum of the gene expression intensities of the channels and the ordinate indicates a difference between them. In areas where the sum of the gene expression intensities of the channels is small or large, the difference of the gene expression intensities between the channels is not so much affected by a true gene expression difference. It is then thought to occur due to a difference in sensitivity between fluorescent dyes of the channels. Thereby, we can observe the distortion thought to occur due to the difference in sensitivity between the fluorescent dyes in Fig. 9.

Referring to Fig. 10, there is shown a diagram of

- 16 -

gene expression intensities of the spot positions of the original data in the channel 1. Referring to Fig. 11, there is shown a diagram of gene expression intensities of the spot positions after the first process in the channel 1.

5 Referring to Fig. 12, there is shown a diagram of gene expression intensities of the spot positions after the second process in the channel 1. They show that the systematic distortion periodically repeated on a grid by grid basis, depending on the spot position, has been  
10 removed by correction.

Referring to Fig. 13, there is shown a diagram of gene expression intensities of the spot positions after the third process in the channel 1. Referring to Fig. 14 to Fig. 17, there are shown diagrams of gene expression  
15 intensities of the spot positions of the original data, after the first process, after the second process, and after the third process in the channel 2. Similarly to the channel 1, they show that the systematic distortion periodically repeated on a grid by grid basis, depending on  
20 the spot position, has been removed by correction.

Referring to Fig. 18 to Fig. 21, there are shown S-D plots of the original data, after the first process, after the second process, and after the third process. As  
25 is shown by Fig. 21, it is apparent that the distortion caused by the difference in sensitivity between the fluorescent dyes has been removed by correction.

According to the present invention, the standardization is performed by combining the standardization with the stable order statistics, 25 and 50

- 17 -

percent points, for fluctuations of the position and scale among the grids (global normalization) and the correction of a distortion depending on the spot position within a grid (local normalization). Thereby, it is possible to simultaneously correct a systematic error caused by deviation of the gene expression intensities among grids or by fluctuations of sensitivity and a distortion depending on the spot position within a grid, almost without any effect of a gene expression frequency nor an outlier.

Furthermore, according to the present invention, a difference in sensitivity between the fluorescent dyes can be easily obtained by using a sum and a difference of the gene expression intensity data in the S-D plot, thus enabling an accurate extraction of a measurement error caused by the difference. Thereby, it is possible to correct efficiently a distortion of measurements caused by the difference in sensitivity between the fluorescent dyes.

Referring to Assumption 2 above,  $L_k(c)$  and  $M_k(c)$  indicate 25 and 50 percent points respectively. Supposing that  $A_k(c)$ ,  $L_k(c)$  and  $M_k(c)$  indicate 35, 10 and 90 percent points of the fluorescence intensity  $y_{ij}^k(c)$  obtained in the channel  $c$  in the grid  $k$  respectively, it is assumed that  $A_k(c)$  and  $M_k(c) - L_k(c)$  are more common to all grids and channels. Thereby, it is possible to effectively correct a systematic error. In this case, the data standardization unit 21 standardizes the expression intensity data it has received by using grid-by-grid order statistics as represented by the following EQ11 (global normalization) (referring to Fig. 2).

- 18 -

$$w_{ij}^k(c) = \frac{y_{ij}^k(c) - A_k(c)}{M_k(c) - L_k(c)},$$

$$c = 1, 2, i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K. \quad (11)$$